基于 CapsNet 的中国手指语识别 *

郝子煜，阿里甫・库尔班[†]，李晓红，依沙・吾阿提别克

(新疆大学 软件学院，乌鲁木齐 830046)

摘　要：中国手指语的识别作为中国手语识别中重要的组成部分，使听障者的交流和人机交互更加便捷。传统的手指语识别采用卷积神经网络的方法，模型结构单一，在池化层会丢弃很多信息。Capsule（胶囊）是在神经网络中构建和抽象出的子网络，每个胶囊都专注于一些单独的任务，又能保留图像的空间特征。分析了中国手语中手指语的特征，构建并扩展了手指语图片训练集，试图用 CapsNet（胶囊网络）模型解决手指语的识别任务，对比了不同参数下 CapsNet 的识别率，并与经典的 GoogLeNet 卷积网络作对比。实验结果表明，CapsNet 在手语识别任务上能达到较好的识别效果。

关键词：手语；手指语识别；神经网络；胶囊网络

中图分类号：TP183　　doi: 10.3969/j.issn.1001-3695.2018.05.0341

Chinese finger language recognition use CapsNet

Hao Ziyu, Alifu.kuerban[†], Li Xiaohong, Yisa.watbek

(*School of Software, University of Urumqi* 830046, *China*)

Abstract: As an important part of Chinese sign language recognition, the recognition of Chinese finger language makes the communication of the deaf and man-machine interaction more convenient. Traditional finger-language recognition adopts the method of convolution neural network(CNN) , leading to the structure of the model is single and a lot of information will be discarded in the pooling layer. Capsules are kinds of constructed and **Abstract:** d subnetworks in neural networks, and meanwhile each Capsule focuses on individual tasks and preserving spatial features of the image. Analyzing characteristics of finger language in Chinese sign language, and constructing and expanding training set of finger language pictures, we try to solve the task of finger language recognition by using CapsNet. Comparing the CapsNet recognition rate under different parameters and comparing with the classic GoogLeNet convolution network, experimental results show that CapsNet can achieve better recognition effect in the task of sign language recognition.

Key words: sign language; finger language recognition; neural network; capsnet

0　引言

手语作为聋哑人思想交流和人际交往的主要工具，在该群体知识习得、个体发展和社会认知中伴有重要的作用[1]。手语是聋哑人通过手和手臂，同时借助头部动作、脸部表情和肢体姿态进行交流的特殊语言。手语识别利用模式识别技术，通过分析手和手臂的动作姿态特征，将序列特征作为分类器的输入进行分类识别任务，最终将手语翻译为文本或声音输出，方便听障人群的日常交流[2]。手语识别也为健全人学习和理解手语提供了便利条件。

手指语用指式代表字母，按照汉语拼音方案拼成普通话。由于手指语指式少，而且易学易记，可以帮助聋哑学生识记、辨认语音，提高看话能力，加快识字进度，更好地掌握新词。手指语作为中国手语重要组成部分，使中国手语更加完善。通过结合手指语改善了手势语的表达方法，使手势语更加精确和丰富。《汉语手指语字母方案》共规定了 30 个字母指式，如图 1 所示。

顾定倩等人[3]认为，用手指语表示特定的手语意义是中国手语中相当普遍的现象。手指语不仅充当某个词素，还经常充当基本词。主要有以下三种形式的手指语：

a)单一字母手势。它是用一个表示声母的手指语来表示一个词的手势，除了“v”这个手指字母没有运用外，其他 29 个手指语全都作为基本词独立出现过，如手势语“碧绿”和“白”，手势如图 2 所示。

收稿日期：2018-05-18；**修回日期**：2018-07-17　　**基金项目**：国家自然科学基金资助项目（61163029）

作者简介：郝子煜（1993-），男，山西吕梁人，硕士，主要研究方向为自然语言处理、图像识别；阿里甫・库尔班（1967-），男（通信作者）（维吾尔族），教授，硕导，主要研究方向为少数民族自然语言处理、虚拟现实技术（Ghalipk@xju.edu.cn）；李晓红（1993-），女，新疆奎屯人，硕士，主要研究方向为图像识别；依沙・吾阿提别克（1992-），男（哈萨克族），新疆伊犁人，硕士，主要研究方向为虚拟现实技术、自然语言处理.

b)字母变式手势。这类基本词的构思仍然是使用手指语，但是在使用方式上发生了变异，称之为变式。例如用一、两个变换方向、或附加动作、或置于身体某个部位的手指语表示一个词的手势，如手势语“冯”。

c)字母音节手势。它是用完整表示声韵母音节手指语表示词的手势，如手势语“吴”。“冯”和“吴”的手势如图 3 所示。



图 1　中国手指语指式图

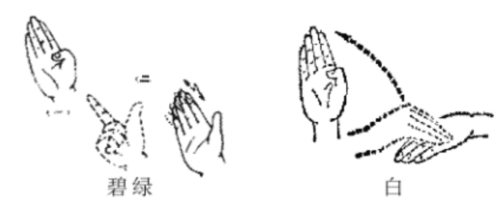


图 2　手语“碧绿”和“白”

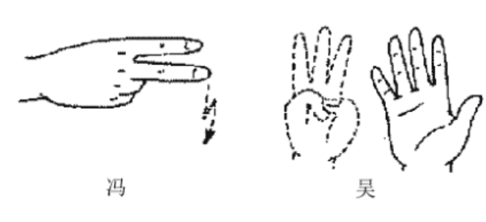


图 3　手语“冯”和“吴”

由此可以看出，手指语是中国手语中必不可少的一部分，不仅可以单独的手语表达，也可以伴随着一、两个手势动作表达手语信息。手指语使手语表达更加准确和丰富，尤其有利于一些生僻字词的表达。

一方面，手语识别可以作为健全人与聋哑人之间的翻译，为聋哑人提供更好的服务；另一方面，作为人体语言理解的一部分，手语识别可作为人机交互的一种手段。综上所述，手指语的识别是手语识别的重要基础和组成部分，对于中国手语的识别具有重大的意义[4]。

1　CapsNet 模型

目前基于计算机视觉的手语识别方法主要采用卷积神经网络的方法[5]。卷积神经网络在图像分类的任务上有突出的表现。一个完整的卷积网络通常包括输入层、卷积层、池化层、全连接层。

传统的卷积神经网络都是在 LeNet5 模型的基础上，用增加网络的层数、改变激活函数等方法改进[6]。LeNet5 是第一个成功应用于数字识别问题的卷积神经网络，其结构如图 4 所示。

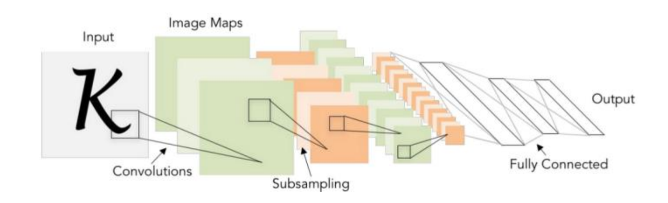


图 4　LeNet5 结构

卷积神经网络通过结合局部感知区域、共享权值、空间或时间上的池化降采样三大特点来充分利用数据本身包含的局部性特征，并且在一定程度上保证位移的不变性[7]。卷积模型的权值共享结构相似于生物神经网络，降低了网络模型的复杂度，减少了权值的数量。因为这种结构特点使其尤其适合大图像数据的机器学习，可以使数量庞大的图像识别问题不断降维，池化结构极大地提高了网络运算的效率。但正是由于这样的池化结构，在对特征图进行抽样时，也会造成丢失一些有效的数据信息。深度学习之父 Hinton 认为，池化解决的问题是错的，我们应该整理信息而不是丢弃信息，并提出了 Capsule 理论[8]。

CapsNet 由 n 个子网络（Capsule）构成，每个胶囊都专注于做一些单独的任务，而胶囊本身需要多层网络来实现。其输出的向量包括物体所属类型的概率以及物体的状态信息（如位置、方向、大小、形变、速率、颜色等）。低层 Capsule 输出的参数会被转换成高层胶囊对实体状态的预测，如果预测一致，则输出这一层的参数。CapsNet 模型如图 5 所示。

可以看出，CapsNet 模型也采用卷积结构提取特征，但 Primary Caps（Capsule 的准备层）可以把数据信息在多通道下分为若干个单元，从而按照每个单元生成保留空间信息的向量，最后输入下一层的 Capsule 神经元中。这一结构取代了传统卷积网络中的池化层，有效地减少了信息的损失。最后一层与全连接层类似，但每个神经元被改造为 Capsule 结构进行分类输出，称为 DigitCaps 层。

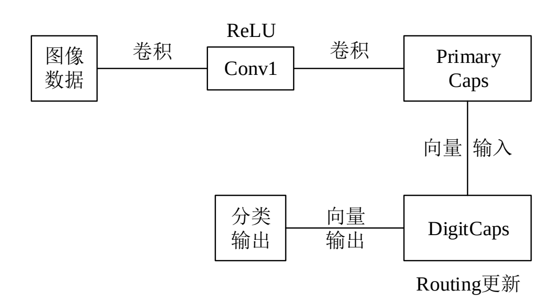


图 5　CapsNet 模型

CapsNet 模型用活动向量表示一个实体是否出现以及这个实体的属性。用向量不同维度上的值分别表示不同的属性，然后用整个向量的模表示这个实体出现的概率。为了保证向量的长度，也就是实体出现的概率在 0~1 间，向量通过一个非线性计算进行压缩和标准化，这样向量在高维空间中的方向体现了这个实体的不同属性。采用 squashing 非线性函数可保证输出向量的长度在 0~1 间。以下是 squashing 函数表达式：

$$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2}\frac{s_j}{\|s_j\|} \tag{1}$$

其中：v 为 Capsule 的输出向量；s 为上一层 Capsule 输出的向量加权和。该非线性函数既保留了输入向量的方向，也将输入向量的长度压缩在[0,1)区间内。向量的输入可分为两个阶段，如下所示：

$$s_j = \sum_i c_{ij}\hat{u}_{j|i} \tag{2}$$

$$\hat{u}_{j|i} = W_{ij}u_i \tag{3}$$

其中：$\hat{u}_{j|i}$ 由较低层的 Capsule 输出 u_i 与权重矩阵相乘得来；c_{ij} 是动态路由过程中的耦合系数，如下所示：

$$c_{ij} = \frac{\exp(b_{ij})}{\Sigma_k \exp(b_{ik})} \tag{4}$$

$$b_{ij} + \hat{u}_{j|i}\cdot v_j \rightarrow b_{ij} \tag{5}$$

利用预测向量 $\hat{u}_{j|i}$ 和输出向量 v_j 的内积来度量向量间的一致性，并更新 b_{ij}，利用 softmax 更新耦合系数，进一步修正下一层 Capsule 的输入 s_j，最后输出新的 v_j。通过这种方式不断迭代更新一致性参数。Capsule 层级间结构如图 6 所示。

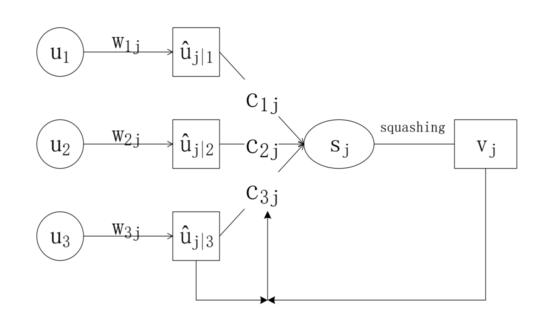


图 6　Capsule 层级间结构

这样通过上述方法不断更新 c_{ij}，不需要应用反向传播算法。而且该 Routing 算法十分容易收敛，通过 3 次迭代就会达到不错的收敛效果。但在整个网络中其他卷积参数和 Capsule 内的 w_{ij} 权重矩阵依然需要根据损失函数进行更新，通常使用标准的反向传播更新这些参数。

2　实验准备

2.1　手指语采集及预处理

Kinect 是微软发布的深度摄像头[9]，提供的深度数据和人体骨骼点数据为手势识别的研究开辟了更广阔的空间。通过 Kinect 中 BodyIndex 方法可以得到深度图像，把手部图像提取出来[10]。

由于 Kinect 的深度值会随距离的增加发生偏差，为达到更好的实验效果，被拍摄的手部位于 Kinect 正前方 1.2 m~1.5 m 的距离，并且分别从手势的正前方、左侧方、右侧方三个角度采集。得到的二值手指语图像将手部置于图像中心，并缩小为 44x44 的二值图像，以减小实验的运算量。

2.2　图像滤波

传统的滤波算法有均值滤波、中值滤波、高斯滤波等。均值滤波算法相对简单且易于实现，但会使目标边缘模糊，而且对 0 值噪声敏感，影响后续处理。高斯滤波处理图像的平滑程度取决于标准差，离中心越近的像素权重越高，平滑效果较好。由于通过 Kinect 采集的深度图像中噪声多为 0 值点，即摄像机无法获取深度值的点，使用中值滤波可以有效去除噪声点，又能保护手部边缘信息[11]。所以本文采用中值滤波算法进行去噪。

中值滤波法是一种非线性平滑技术，它的基本原理是通过使用模板合算子对覆盖区域内所有像素值排序，将这些像素点的中值更新当前像素点的值。本文使用的中值滤波模板大小为 3x3。

2.3　扩展数据集

在训练图像识别的深度神经网络时，通过使用大量的训练数据，使网络得到更好的性能，如提高网络的分类准确率、防止过拟合等。获取更多的训练样本的代价很大，在实践中常常是很难达到的，但是通过人为扩展训练数据[12]能够获得类似的效果。

本文对三个方向的手指语图像采用水平翻转、旋转、添加随机椒盐噪声方法来扩展训练数据集，这些方法可以模拟真实世界的变化，提升模型的准确率和泛化能力[13]。

采集中国手指语前 10 个字母 a~j 的手指语，对数据集全部处理完毕后，共生成 6 500 张手指语图片，其中随机选取 500 张作为测试集，数据处理流程如图 7 所示。

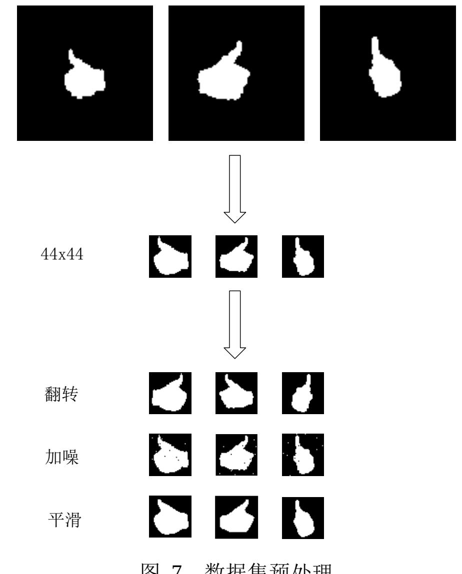


图 7　数据集预处理

3　实验

实验硬件环境为：Intel 酷睿 i5-3230 CPU @ 2.60 GHz 四核 CPU，8 GB 内存，NVIDIA GeForce GT 645M 2 GB 显卡。首先输入 44x44 图像数据。由于手指语类型较多，手形轮廓较复杂，

较大的卷积核可以提取到较多手型边缘特征。本实验以前两层卷积核设置大小为 8x8、10x10 和 13x13，并在不同环境下作为对比实验[14]，第一层卷积步幅为 1x1，第二层卷积步幅为 2x2，反向传播算法迭代 20 次，动态路由算法共迭代 3 次。

首先卷积核大小设置为 10x10，对全部图像进行中值滤波处理，在没有添加噪声的条件下进行了实验，达到了较高的识别率。损失函数变化如图 8 所示。由图可以看出，CapsNet 模型可以很快地拟合数据。在训练到第 4 500 步后，趋于稳定状态。训练效果如图 9 所示。在该实验条件下，CapsNet 模型在测试集的识别率达到 95.8%。

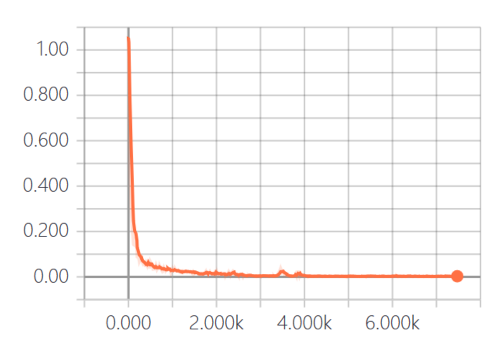


图 8　整体损失变化

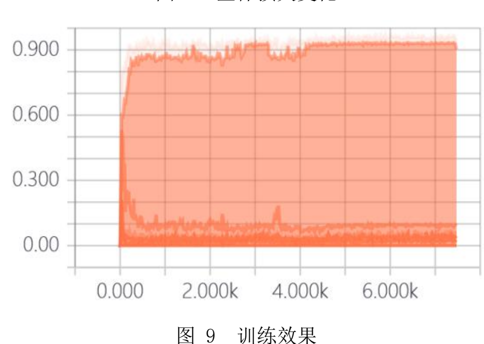


图 9　训练效果

在上述实验的基础上添加随机椒盐噪声后，识别率下降并不明显，证明了此模型拥有较好的鲁棒性。本文分别使用 8x8 和 13x13 的卷积核重复实验，实验结果对比如表 1 所示。

表 1　不同参数下平均识别率对比

	8x8	10x10	13x13
无噪声	92.6%	95.8%	95.8%
有噪声	91.2%	94%	95.4%

在此基础上，实验分别在测试集上验证每个手指语的准确率，达到了预期效果。由于手指语的采集分别从正前方、左侧方、右侧方三个角度采集，导致手指语 g 和 i 中几个手势相似，所以在测试集上的识别率不太理想，平均识别率分别达到 84% 和 82%，而手指语 j 的识别率达到 100%。单个手指语识别率如图 10 所示。

GoogLeNet 的 Inception 结构相比 AlexNet 和 VGG，增加了网络宽度，拥有更少的参数，保持了网络结构的稀疏性，利用密集矩阵极大的提高了计算性能[15]。Inception –v4 [16]不仅具有 Inception 前四个版本的特性，更是结合了 ResNet，进一步减小了错误率。由此，在有噪声的条件下，使用 Inception –v4 进行实验作为对比。CapsNet 与 Inception –v4 实验对比如表 2 所示。从表 2 可以看出，Inception –v4 的平均准确率为 94.4%，略低于 CapsNet。对于单个手指语字母的识别，最高都达到了 100%。

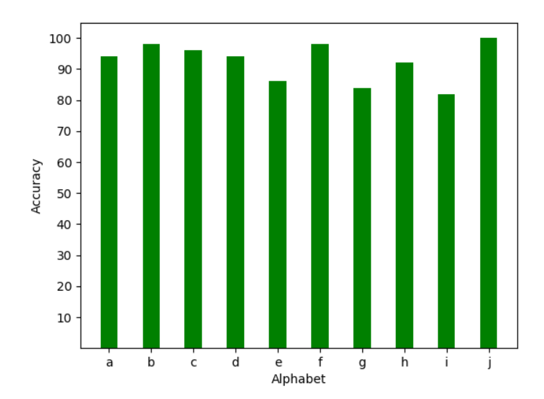


图 10　单个手指语准确率

表 2　CapsNet 与 Inception-4 实验对比

网络结构	平均准确率	单个最高准确率
CapsNet	95.4%	100%
Inception –v4	94.4%	100%

4　结束语

本文采用 CapsNet 模型，使用卷积结构和动态路由更新参数的算法进行静态手指语的识别。实验显示 CapsNet 模型在手指语识别的任务上达到很好的效果，在添加噪声的条件下最高平均识别率达到 95.4%，尤其是在三个角度的手指语会造成部分手指语特征不明显的情况下达到较高的准确率，说明了这种动态路由更新参数和用向量预测目标属性的算法有较好的性能。手语包括了手形、方向、位置、运动轨迹四个特征。手指语只是手形的一部分。在未来的研究工作中，针对动态手语的多个手语特征，使用时空域的算法结合 Capsule 的优良特性进行完整手语的识别。

参考文献：

[1] 余晓婷，贺荟中. 国内手语研究综述 [J]. 中国特殊教育，2009 (4): 36-41. (Yu Xiaoting, He Huizhong. A review on domestic sign language study [J]. Chinese Journal of Special Education, 2009 (4): 36-41.)

[2] 姚登峰，江铭虎，阿布都克力木・阿布力孜，等. 中国手语信息处理述评 [J]. 中文信息学报，2015, 29 (5): 216-228. (Yao Dengfeng, Jiang Minghu, Abudoukelimu Abulizi, *et al.* A survey of Chinese sign language processing [J]. Journal of Chinese Information Processing, 2015, 29 (5): 216-228.)

[3] 顾定倩，宋晓华，于缘缘. 中国手语基本词（基本动作）类型分析 [J]. 中国特殊教育，2005 (2): 65-72. (Gu Dingqian, Song Xiaohua, Yu Yuanyuan. The analysis of Chinese sign language's basic words (basic movements) [J]. Chinese Journal of Special Education, 2005 (2): 65-72.)

[4] 李勇，高文，姚鸿勋. 基于颜色手套的中国手指语字母的动静态识别 [J]. 计算机工程与应用，2002, 38 (17): 55-58. (Li Yong, Gao Wen, Yao Hongxun. Chinese sign language finger alphabet recognition based on

color gloves [J]. Computer Engineering & Applications, 2002, 38 (17): 55-58.)

[5] 易靖国，程江华，库锡树. 视觉手势识别综述 [J]. 计算机科学，2016, 43 (s1): 103-108. (Yi Jingguo, Cheng Jianghua, Ku Xishu. Review of gestures recognition based on vision [J]. Computer Science, 2016, 43 (s1): 103-108.)

[6] 周飞燕，金林鹏，董军. 卷积神经网络研究综述 [J]. 计算机学报，2017, 40 (6): 1229-1251. (Zhou Feiyan, Jin Linpeng, Dong Jun. Review of convolutional neural network [J]. Chinese Journal of Computers, 2017, 40 (6): 1229-1251.)

[7] 李旭冬，叶茂，李涛. 基于卷积神经网络的目标检测研究综述 [J]. 计算机应用研究, 2017, 34 (10): 2881-2886, 2891. (Li Xudong, Ye Mao, Li Tao. Review of object detection based on convolutional neural networks [J]. Application Research of Computers, 2017, 34 (10): 2881-2886, 2891.)

[8] Sabour S, Frosst N, Hinton G E, *et al.* Dynamic routing between capsules [C]// Advances in Neural Information Processing Systems. 2017: 3856-3866.

[9] 刘佳，郑勇，张小瑞，等. 基于Kinect的手势跟踪概述 [J]. 计算机应用研究, 2015, 32 (7): 1921-1925. (Liu Jia, Zheng Yong, Zhang Xiaorui, *et al.* Overview of hand gesture tracking based on Kinect [J]. Application Research of Computers, 2015, 32 (7): 1921-1925.)

[10] 邓瑞，周玲玲，应忍冬. 基于 Kinect 深度信息的手势提取与识别研究 [J]. 计算机应用研究, 2013, 30 (4): 1263-1265. (Deng Rui, Zhou Lingling, Ying Rendong. Gesture extraction and recognition research based on Kinect depth data [J]. Application Research of Computers, 2013, 30 (4): 1263-1265.)

[11] 阮秋琦. 数字图像处理学 [M]. 北京：电子工业出版社, 2007. (Ruan Qiuqi. Digital image processing [M]. Beijing: Electronic Industry Press, 2007.)

[12] Krizhevsky A, Sutskever I, Hinton G E, *et al.* Imagenet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. 2012: 1097-1105.

[13] Simard P Y, Steinkraus D W, Platt J, *et al.* Best practices for convolutional neural networks applied to visual document analysis [C]// Proc of International Conference on Document Analysis and Recognition. 2003: 958-963.

[14] Szegedy C, Liu Wei, Jia Yangqing, *et al.* Going deeper with convolutions [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. I.] : IEEE Press, 2015: 1-9.

[15] Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications [J/OL]. [2018-03-07]. https://arxiv. org/abs/1605. 07678.

[16] Szegedy C, Ioffe S, Vanhoucke V, *et al.* Inception-v4, Inception-ResNet and the impact of residual connections on learning [C]// Proc of the 31st AAAI Conference on Artificial Intelligence. Palo Alto, California: AAAI Press, 2017: 4278-4284.